Tribhuvan University

Institute of Science and Technology

A Project Report

On

Breast Cancer Detection using Logistic Regression

Submitted to

Department of Computer Science and Information Technology

Swastik College, Chardobato, Bhaktapur

In partial fulfilment of the requirements for the Bachelor's Degree in Computer Science And Information Technology

Submitted by

Bishesh Neupane (24367/076)

Dev Raj Basnet(24370/076)

Ram Timilsina(24383076)

March 2024

Tribhuvan University

Institute of Science and Technology

A Project Report

On

Breast Cancer Detection using Logistic Regression

Submitted to

Department of Computer Science and Information Technology

Swastik College, Chardobato, Bhaktapur

In partial fulfilment of the requirements for the Bachelor's Degree in Computer Science And Information Technology

Submitted by

Bishesh Neupane (24367/076)

Dev Raj Basnet(24370/076)

Ram Timilsina(24383076)

March 2024

Students Disclaimer

We hereby declare that this study entitled “Breast Cancer Detection using Logistic Regression” is based on our original research work. Related works on the topic by other researchers have been duly acknowledged. We owe all the liabilities relating to the accuracy and authenticity of the data or any information included here under.

……. ……..

Bishesh Neupane Dev Raj Basnet

……..

Ram Timilsina

Supervisor's Recommendation

I hereby recommend that this project report prepared under my supervision by the team of Bishesh Neupane, Dev Raj Basnet and Ram Timilsina entitled "Breast Cancer Detection using Logistic Regression" in partial fulfillment of the requirements for the degree of B.Sc in Computer Science and Information Technology to processed for the evaluation.

…….

Mr. Sagar Rana Magar

Lecturer, Swastik College

Letter of Approval

We certify that we have read this report and in our opinion, it is satisfactory in the scope and quality as a project report in the partial fulfillment to the requirement for the Bachelor's Degree in Computer Science and Information Technology.

…… Mr. Sagar Rana Magar Supervisior Lecturer,Swastik College	…… Mrs. Shristi Khatiwoda Head of Department Swastik College
…… Internal Examiner	…… External Examiner

Acknowledgement

This report titled "Breast Cancer Detection using Logistic Regression," has been crafted in fulfillment of the undergraduate project requirement. We extend our sincere gratitude to Tribhuvan University for integrating a project report into the bachelor's program, providing an invaluable opportunity to apply practical knowledge to any chosen topic. This Undergraduate Project serves as a platform for gaining practical insights into the chosen area of interest and refining professional skills, bridging the gap between academic learning and practical application.

We extend our heartfelt thanks to Mrs. Shristi Khatiwoda, Head of the Department and Mr. Sagar Rana Magar, Supervisor both from the Department of Computer Science and Information Technology, for their unwavering support and for granting us the opportunity to embark on this project.

Thanking You,
Bishesh Neupane
Dev Raj Basnet
Ram Timilsina

Abstract

Breast cancer is one of the most prevalent and life-threatening diseases affecting women worldwide. Early detection is crucial for improving the prognosis and increasing the chances of successful treatment. This project presents the state of the art breast cancer detection system based on deep learning techniques.

The system is constructed using the python with Django ,Data Science using Regression Analysis (Logistic regression). The system is trained on large and diverse dataset to ensure robust and accurate performance. The breast cancer detection system involves testing on a separate dataset to access its overall accuracy. The result demonstrate the system ability to provide reliable and fast detection of potential breast cancer case, enabling intervention and reducing false positives.

Further more ,the system is designed to be user friendly and health professionals to integrate it into their existing workflows. It potentials to assists medical experts in their decision-making process makes it valuable tool for early breast cancer detection and contributing to improved patient outcomes and quality of life.

Keywords: Breast Cancer (BC), Django , Logistic Regression (LR), early diagnosis

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AUC Area Under Curve

BC Breast Cancer

LR Logistic Regression

ML Machine Learning

Chapter 1: Introduction

1.1 Background

The field of health care has incorporated various technologies such as managing and accessing health details of patients and related advises. The detection of cancer has been historically difficult when it's comes to determining the best course of action for haematological diseases. A considerable portion of the population experiences one or multiple diseases, and while medical science has made significant strides, there remains a significant percentage of the population afflicted with health problems that may be life-threatening.To improve the accuracy in detecting fatal conditions promptly, employing safe and practical methods and utilizing latest technical methodologies can help in reducing the effect of the diseases and decreases healthcare expenses. When the atypical cells of the body begin to divide and infiltrate healthy cells, it results in cancer. Globally, Breast Cancer (BC) threatens more women compared to any other type of cancer. It is responsible for the majority of newly diagnosed cancer cases and cancer-related fatalities, which makes it a major public health concern in modern society.

This study is focused on predicting the breast cancer using Logistic Regression (LR) and the model is further elaborated with LR algorithms. Logistic Regression serves as a vital statistical tool for binary classification tasks. The project typically involves predicting whether a given breast tumor is malignant or benign based on various medical data features, such as radius, smoothness, Concavity etc. The dataset used in the project comprises labeled examples, each with input features and corresponding class labels denoting malignancy or benignancy. Logistic regression models the probability of a malignant tumor using a logistic (sigmoid) function, mapping the linear combination of input features to a probability value between 0 and 1. The performance is evaluated on a separate test set and the area under the ROC curve. The coefficients of the logistic regression model provide insights into the influence of each feature on the probability of malignancy. Ultimately, a successfully trained logistic regression model can be deployed for real-world breast cancer detection applications.

1.2 Problem Statement

Breast Cancer is a complex health problem and we often find breast cancer too late, making treatment harder. The existing challenges in breast cancer detection lead to several consequences. Delayed diagnoses result in advanced stages of cancer, negatively impacting treatment outcomes. Disagreements among pathologists contribute to confusion and the potential for misdiagnoses. The inefficiencies in manual detection systems, marked by time constraints and human errors, compromise the overall effectiveness of the diagnostic process.

To overcome these challenges, the implementation of a computer-aided diagnosis system proves instrumental in eliminating issues associated with breast cancer detection. This system significantly enhances accuracy, providing results in a much shorter timeframe. So Computer based breast cancer detection system using Logistic Regression (LR) is employed and adding the smart and efficient touch to this project.

1.3 Objective

The main objective of this project is to develop a Breast cancer detection system using logistic Regression.

1.4 Scope and limitation

1.4.1 Scope

The project is dedicated to the development of a comprehensive system designed for the early detection of breast cancer. This system is imagined to be implemented in diverse healthcare settings, ranging from hospitals and screening programs to online medical platforms. Its primary objective extends beyond aiding medical professionals in identifying breast cancer at its initial stages. It also seeks to contribute significantly to research endeavors by analyzing extensive datasets. Through this dual approach, the project aims to enhance existing diagnostic tools and methodologies for breast cancer. The goal is to make a real impact on how we deal with breast cancer, making it easier to detect and improving our understanding of it.

1.4.2 Limitation

The Breast Cancer Detection System project can’t give the result by taking X-ray report image as input. If the model is too complex compared to the dataset, it might learn too much from the training data and struggle with new information. The success of the model depends on having

good and complete data so mistakes or missing information can make it less reliable. Selecting the right data for the model is tricky. If we miss important details or include irrelevant ones, it can affect how well the model works. If there's a significant imbalance in the number of different types of tumors, the model might become biased.

1.5 Development Methodology

The model chosen for this system is “Iterative Waterfall model”. The iterative waterfall model combines elements of both the waterfall model and iterative development. The iterative waterfall model allows for better communication between team members, better risk management and improved project control. It also provide more flexibility in making changes to the project scope as each phase can be revisited and revised. This is a simple project with the well-defined process and the requirement. The various step of the iterative waterfall model are followed throughout the whole project. This model is understood and incorporate in these project. Thus, the system is developed according to the iterative waterfall model. The various steps of this model is shown in figure below.

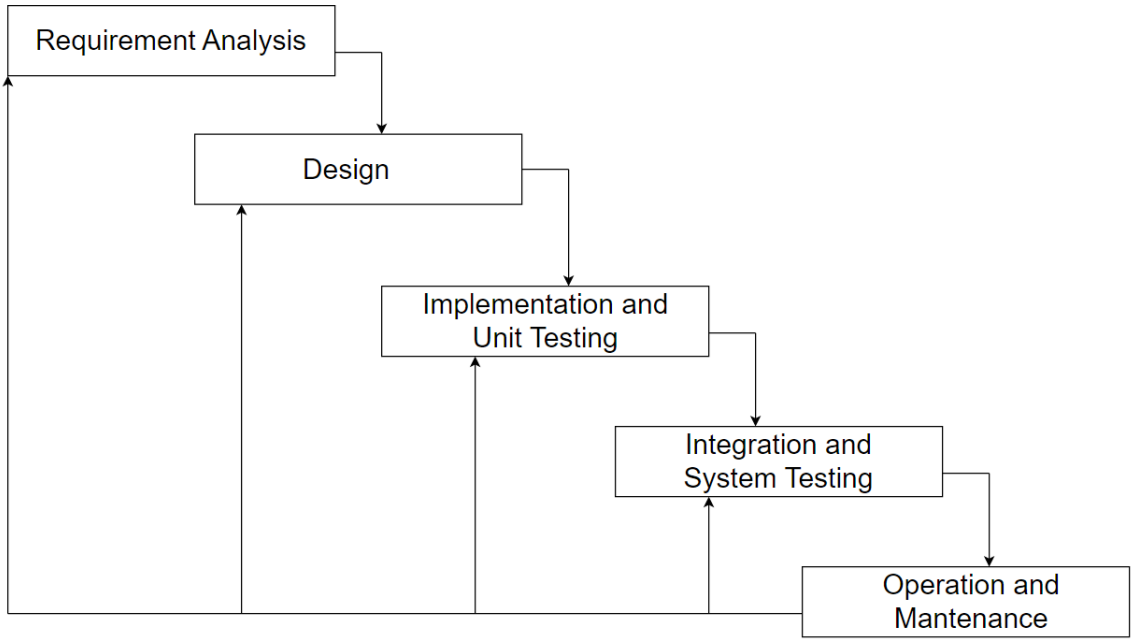


Figure 1.1: Iterative Waterfall Development Model

1.6 Report Organization

The report is divided into six different chapters which are as follows:

The report begins with **Chapter 1** that explains about introduction, problem statement, objectives, scope and limitations with development methodology. The next **Chapter 2** explains about the background study and literature review related to the paper research for the system development.Then, **Chapter 3** discussed about the system analysis where requirement analysis and feasibility analysis are required in system development. Then, **Chapter 4** deals with the project design and modeling are the core part of the development process. Then, **Chapter 5** discussed about the implementation and testing of the system. Lastly, **Chapter 6** conclude and give the future recommendations on the system development.

Chapter 2: Background Study and Literature Review

2.1 Background Study

In medical research and healthcare, machine learning has emerged as a promising strategy, for early detection and diagnosis of cancer. Machine Learning algorithm find the patterns and connections in massive volumes of patient data that may not be obvious to human physicians. ML algorithms may examine mammography pictures to find minute alterations or anomalies that might be signs of breast cancer. Also, they can help with the interpretation of biopsy data and the search for possible biomakers for certain breast cancer subtypes. There are various shortcomings for this procedure as it involves intra-observer variation, cancer cells and tissues can also have multiple appearances, and many other figures in cells have the same hyperchromatic features, which make identification difficult. The choice of the area is also a factor as the process is done only on a small area of tissue, so the chosen area should be in the tumor periphery. Logistic Regression (LR) is used to solve the classification problems. These features can include various attributes such as radius, texture, smoothness, concavity etc are the characterstic of malignant or benign tumors. Breast Cancer Detection using LR will provide better detection for breast cancer cases from different dataset. This will reduce the chances of mistakes during the diagnosis of either a malignant or benign tumor.

Two different types tumors can be encounter early, those are:

a.Benign Tumor

b.Malignant Tumor

A tumor is an abnormal collection of cells. It forms when cells multiply more than they should or when cells don’t die when they should. A tumor can be malignant or benign.

a.Benign Tumor

Benign tumors are those that stay in their primary location without invading other sites of the body. They do not spread to local structures or to distant parts of the body. Benign tumors tend to grow slowly and have distinct borders. Benign tumors are not usually problematic. However, they can become large and compress structures nearby, causing pain or other medical complications. For example, a large benign lung tumor could compress the trachea and cause difficulty in breathing. This would warrant urgent surgical removal. Benign tumors are unlikely to recur once removed.

b.Malignant Tumor

Malignant tumors have cells that grow uncontrollably and spread locally and/or to distant sites. Malignant tumors are cancerous (i.e, they invade other sites). They spread to distant sites via the bloodstream or the lymphatic system. This spread is called metastasis. Metastasis can occur anywhere in the body and most commonly is found in the liver, lungs, brain and bine. Malignant tumors can spread rapidly and require treatment to avoid spread. If they are caught early, treatment is likely to be surgery with possible chemotheraphy or radiotheraphy. If the cancer has spread, the treatment is likely to be systemic, such as chemotheraphy or immunotheraphy.

2.1.1 Logistic Regression Analysis

Logistic regression is supervised learning algorithm which is used to solve the classification problems. In classification problems, we have dependent variables in a binary or discrete format such as 0 or 1. Supervised learning algorithm is a form of machine learning in which the algorithm is trained on labeled data to make predictions or decisions based on the data inputs. Logistic regression algorithm works with the categorical variable such as 0 or 1, Yes or No, True or False, Spam or not spam, etc. It is a predictive analysis algorithm which works on the concept of probability. Logistic regression is a type of regression, but it is different from the linear regression algorithm in the term how they are used. Logistic regression uses sigmoid function or logistic function which is a complex cost function. This sigmoid function is used to model the data in logistic regression. Logistic regression is employed in this study is to categorize observations based on a

variety of data sources and to swiftly determine the most effective factors for classification. Figure 2 depicts the logistic function.

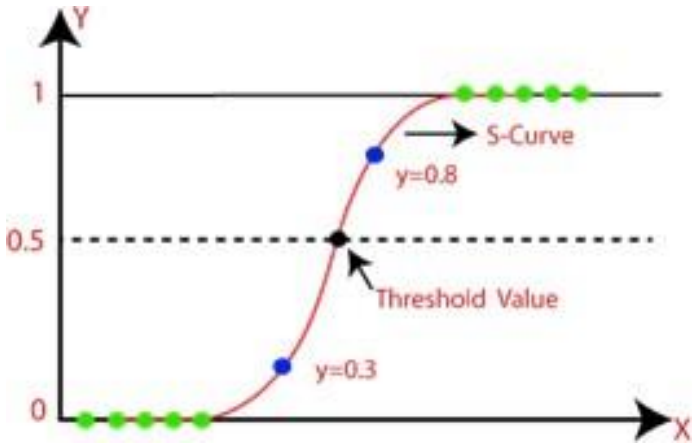


Figure 2.1: Logistic Regression

(a) Decision Boundary

To forecast which class a piece of data belongs to, a threshold might be set.Based on this threshold, the calculated probability is divided into groups. Here,

if (*predicted value*) >=0.5

classify tumors as malignant else benign.

2.1.2 Sigmoid Function

A mathematical function with a characterstic "S"-shaped curve, also known as a sigmoid curve, is sigmoid function. As sigmoid function translate the entire number line into 0 and 1, or -1 and 1, one of its use is to convert an actual value into one that may be analyzed as a probability. As sigmoid function converges faster than any other activation function, for probabilistic classification, it was selected. As standard choice for a sigmoid function is logistic function. Overall, the sigmoid function is defined as :

$$S(x) = \frac{1}{1 + e^{x}}$$

Where, S(x) = output between the 0 and 1 value.

X = input to the function

e=base of natural logarithm

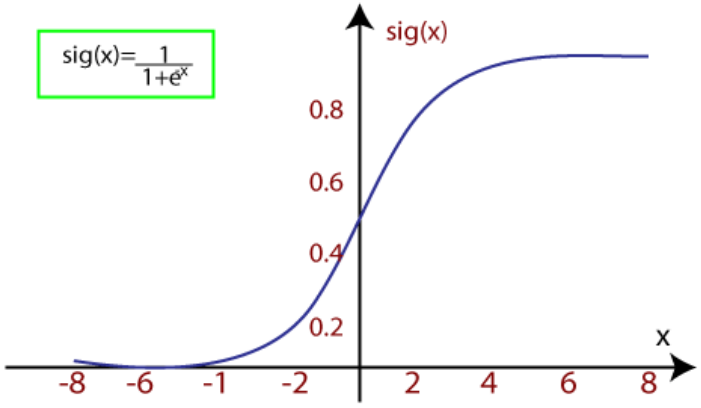


Figure 2.2: Sigmoid Function

2.2 Literature Review

There are many studies done on breast cancer datasets, and most of them have sufficient classification accuracy. Aruna et al. used naïve Bayes, Support vector machine(SVM), and decision trees to classify a Wisconsin breast cancer dataset and got the best result by using SVM with an accuracy score of 96.99%[1].

Chaurasia et al. compared the performance of supervised learning classifiers by using Wisconsin breast cancer dataset and naïve Bayes, SVM, neural networks, decision tree methods applied. According to the study results, SVM gave the most accurate result with a score of 96.84%[2].

Arsi et al. also used the Wisconsin breast cancer dataset and made a performance comparison among machine learning algorithms : SVM, decision tree (c4.5), naïve bayes, and k-nearest neighbors. The study aimed to classify data in terms of efficiency and effectiveness by comparing the accuracy, precision, sensitivity and specificity of each algorithm. The experimental result showed that SVM had the best score with an accuracy of 97.13%[3].

Bernal et al. used clinical data on medical intensive care units. Machine learning techniques such as logistic regression (LR), neural networks, decision trees and k –nearest neighbors were applied to predict the decrease of patient inside the hospital over 24 hours. The highest accuracy scores were obtained with logistic regression (LR) and k-nearest neighbor(KNN-5) technique among training data. Bernal pointed out that it is necessary to decide parameters rather than the algorithm to get better accuracy results[4].

Chapter 3: System Analysis

3.1 System Analysis

System analysis involves requirement analysis: Functional Requirements (Illustrated using use case diagram/use case descriptions), Non-Functional Requirements and Feasibility Analysis.

3.1.1 Requirement Analysis

The system requirement collection specification of the project consists of functional and nonfunctional requirement.

i. Functional Requirements

First the user must provide the attributes and symptoms they have faced then the system predicts the most probable diseases further process is explained by the use case diagram.

a) The System must provide the user/admin to input the data that later feed to algorithm.
b) A Logistic Regression (LR) model must be utilized and trained on the kaggle dataset. The model must take the data as input and predicts result on the basis of trained dataset.
c) The webpage must be designed using the Django framework in Python. Django must be employed to build the user interface and manage user interactions, including data input.
d) Once data is given, the model must process the information and provide a prediction based on the trained dataset. The prediction must be displayed on the webpage to update the UI with the result.
e) Users must be able to view the results through the system, showcasing predictions of tumor types as either benign or malignant.
f) In this system, administrators must have the ability to add datasets, train them, and perform testing as needed.

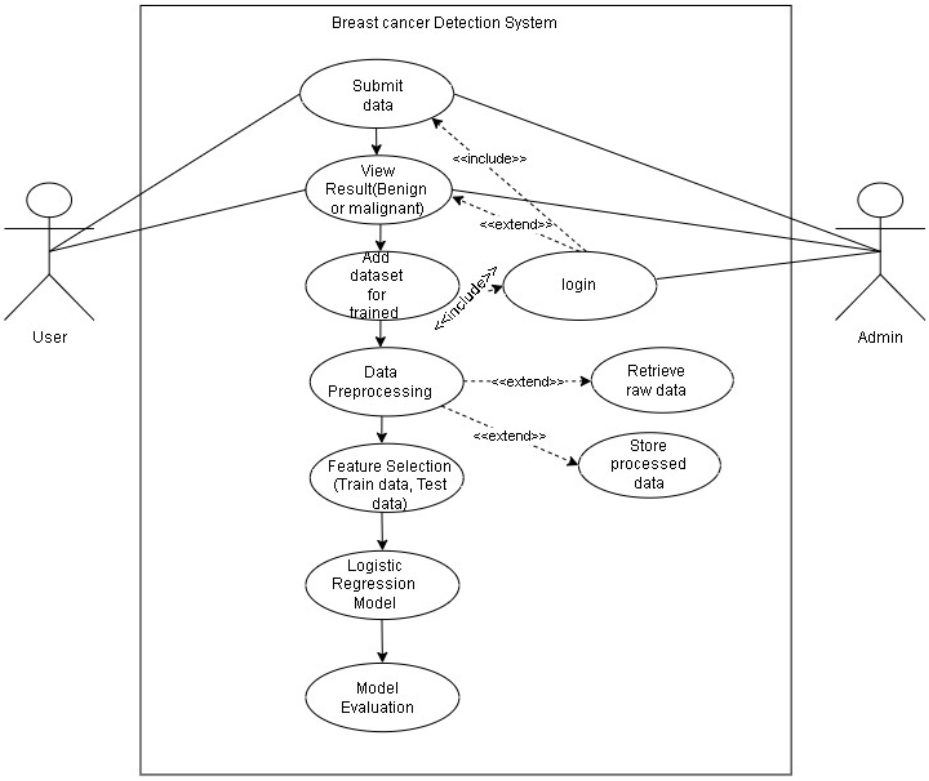


Figure 3.1: Use Case Diagram of Breast Cancer Detection System

ii. Non- Functional Requirements

Some of the non-functional requirements of Breast Cancer Detection System are summarized as following:

a) The system must have high level of accuracy in detecting breast cancer to reduce the number of false positive and false negatives.
b) The system must process data quickly to avoid delays in diagnosis and treatment.
c) The system must be able to handle a large volume of data and should be scalable to handle an increase in the number of users.
d) The system must be reliable, and the results of the detection should be consistent and repeatable.
e) The system must be easy to use and understand for medical professionals, who may not have extensive technical knowledge.
f) The system must be compatible with a variety of devices and operating systems to ensure widespread adoption.

g) The system must be easy to maintain and update, as new data and techniques become available for improving breast cancer detection.
h) The system must be accessed by anyone and anywhere having PC and internet connection.

3.1.2 Feasibility Study

Here, we have studied all the feasibility aspects of the project under consideration to check out if the project is feasible with the decided requirements and availability of information, technologies and budgets.

i. Technical Feasibility

The technical feasibility of the project is contingent on a thorough evaluation of the required technical resources including hardware, software, and other technological components. The following technical requirements are identified for the successful implementation of the breast cancer detection system using regression analysis:

Hardware Requirements

a) Processor: A processor with a dual-core capacity or higher.
b) RAM: A minimum of 1GB RAM to support the computational demands of the system.
c) Storage: At least 15GB of hard disk space for storing datasets, model parameters, and application files.
d) Input Devices: Standard input devices such as a keyboard and mouse for user interactions.

Software Requirements

a) Operating System: Windows 11 for compatibility with the chosen software stack.
b) Integrated Development Environment (IDE): Visual Studio Code to facilitate efficient coding and development.
c) Programming Language: Python for implementing regression analysis and model training.
d) Framework: Django, a Python web framework, for building the user interface and handling interactions.

The selected tools including Visual Studio Code, Python, and Django, are known for their straightforward installation processes. Python with Django is recognized for its efficiency and demands fewer system resources, ensuring optimal performance on the specified hardware. The system's resource requirements align with the designated hardware capacities, avoiding strain. The combination of Windows 11, Visual Studio Code, and Django facilitates the development of user-friendly interfaces for end-users. The technical feasibility of the breast cancer detection system is substantiated by the careful consideration and fulfillment of hardware, software, and additional technical requirements, contributing to a robust and efficient solution.

ii. Economic Feasibility

The development cost for the breast cancer detection system, considering both hardware and software expenses, is considered reasonable. The specified hardware requirements, including are widely available and cost-effective in the current market. These components contribute to the affordability of the overall system setup.

Moreover, the reliance on open-source tools such as Python, Django, and Visual Studio Code helps in mitigating software licensing expenses. The use of these tools allows for efficient development without incurring significant costs related to proprietary software. Open-source solutions are known for their accessibility and cost-effectiveness, contributing to the economic feasibility of the project. The combination of readily available and cost-effective hardware components, along with the utilization of open-source software tools, positions the project as financially viable.

iii. Schedule Feasibility

For each task of the project, proper estimation and splitting of the time have been done. Overall, the calculated time is sufficient enough to complete project in time.

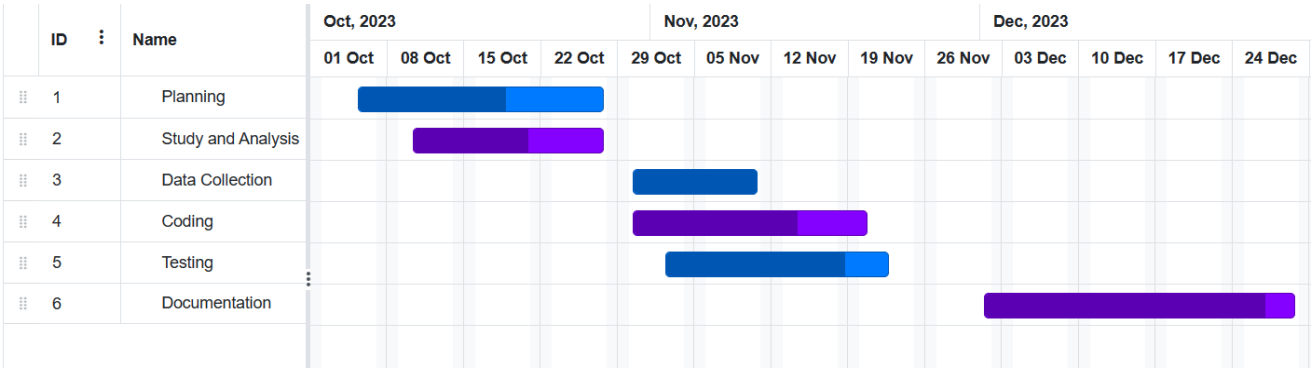


Figure 3.2: Working Schedule

3.1.3 Dataset Description

In this study, publicly available categorical dataset is used which can be taken from kaggle. The datasets are annotated into two categories for diagnosis form: malignant (cancerous) or benign (non-cancerous).

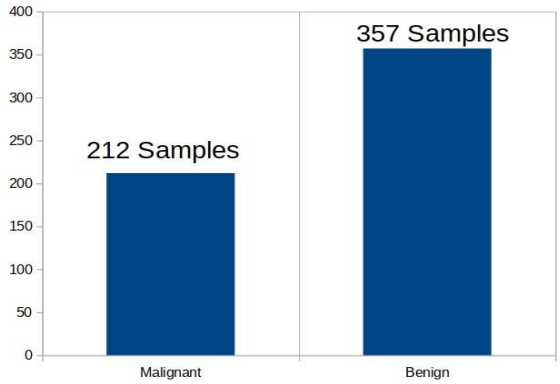


Figure 3.3: Sample Distribution as per the target column

Out of total 569 samples, Figure 5 shows the classification of target column ‘diagnosis’. Among the different 30 features, table 2 shows 10 features for 5 rows. The dataset contain diagnosis as target features while rest other features are training sets. 30 different categories includes ‘radius-mean’,’texture-mean’, ‘perimeter-mean’, ‘area-mean’, ‘smoothness-mean’, ‘compactness-mean’, ‘concavity-mean-concave’, ‘points-mean’, ‘symmetry-mean’, ‘fractal-dimension-mean’, ‘radius-se’, ‘texture-se’, ‘perimeter-se’, ‘area-se’, ‘smoothness-se’, ‘compactness-se’, ‘concavity-se’, ‘concave-point-se’, ‘symmetry-se’, ‘fractal-dimension-mean’, ‘radius-worst’, ‘texture-worst’, ‘perimeter-worst’, ‘area-worst’, ‘smoothness-worst’, ‘compactness-worst’, ‘concavity-worst’, ‘concave-point-worst’, ‘symmetry-worst’, ‘fractal-dimension-worst’.

id	diagnosis	radius_me	texture_m	perimeter	area_mea	smoothne	compactn	concavity	concave p	symmetry	fractal_dir	radius_se	texture_se	perimeter	area_se	smoothne	compactn	concavity	concave p	symmetry	fractal_dir	radius_wo
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193	25.38
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532	24.99
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571	23.57
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208	14.91
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115	22.54
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02165	0.005082	15.47
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01369	0.002179	22.88
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01486	0.005412	17.06
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553	0.01226	0.02143	0.003749	15.49
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01789	0.01008	15.09
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0146	0.003042	19.19
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02008	0.004144	20.42
846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889	0.0409	0.04484	0.01284	20.96
846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02981	0.003002	16.84
84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01961	0.008093	15.03
84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741	0.0109	0.01857	0.005466	17.46
848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998	0.01109	0.0141	0.002085	19.07
84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188	0.01297	0.01689	0.004142	20.96
849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391	0.01521	0.01356	0.001997	27.32
8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462	0.0146	0.02387	0.01315	0.0198	0.0023	15.11
8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.004097	0.01898	0.01698	0.00649	0.01678	0.002425	14.5
8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.009606	0.01432	0.01985	0.01421	0.02027	0.002968	10.23
8511133	M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.006789	0.05328	0.06446	0.02252	0.03672	0.004394	18.07
851509	M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.004728	0.01259	0.01715	0.01038	0.01083	0.001987	29.17

Figure 3.4:Sample Dataset

3.1.4 Analysis

i.Sequence Diagram

The first activity is entering patient data which represents the user's interaction with the system, where the user enters relevant data about a patient. After entering the patient data, the user initiates the process by submitting the data for cancer prediction.The backend of the system takes over and processes the submitted data.The system preprocesses the input data, which involves cleaning the data, handling missing values, normalizing or scaling features, and potentially selecting relevant features for the prediction task. The preprocessed data is used to train the logistic regression model. During this activity, the model's parameters are optimized to make accurate predictions based on the training data. Once the logistic regression model is trained, it is saved for future use. And the saved logistic regression model is used to predict the likelihood of cancer based on the input patient data.Finally,the prediction result is displayed to the user.

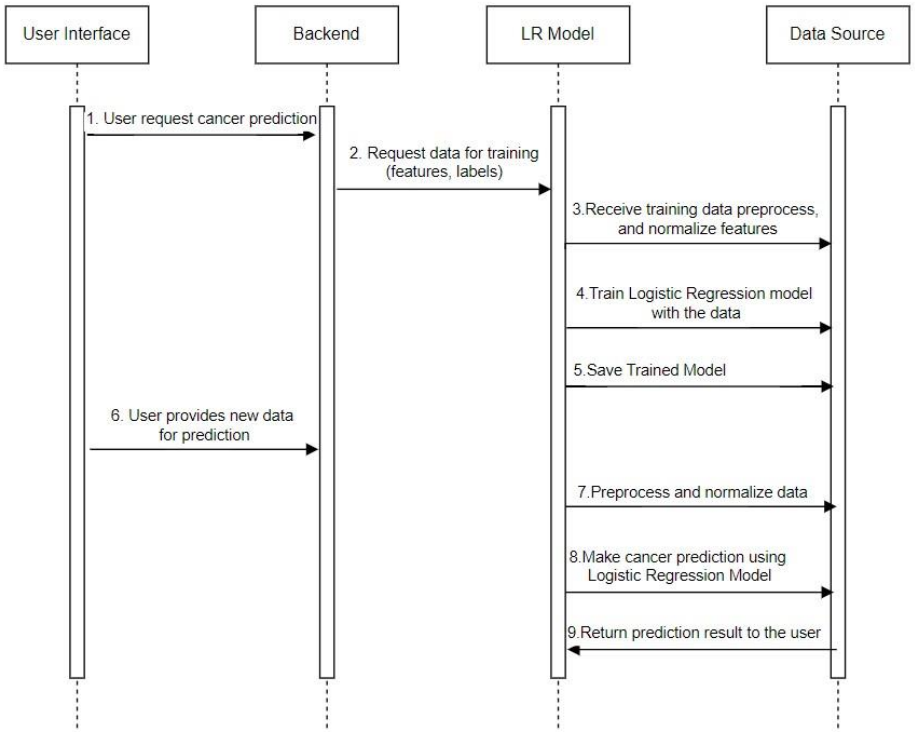


Figure 3.5: Sequence diagram for Breast Cancer Detection System

ii. Activity Diagram

Gather relevant data for cancer prediction which may include patient records, medical history, test results, and other relevant features.Clean and preprocess the collected data which involves handling missing values, normalizing or scaling features, and possibly encoding categorical variables.Divide the dataset into training and testing sets. The training set is used to train the logistic regression model, while the testing set is used to evaluate its performance.Choose relevant features for training the logistic regression model. Feature selection aims to use the most informative features while discarding irrelevant or redundant ones.Train the logistic regression model using the training dataset. This involves optimizing the model's parameters to make accurate predictions.Evaluate the performance of the trained logistic regression model using the testing dataset.Deploy the trained logistic regression model for making predictions on new, unseen data. Users interact with the system, providing new data for cancer prediction.Collect input data from users, ensuring it aligns with the format expected by the trained logistic regression model.Use the trained logistic regression model to make predictions on the new data.Present the prediction results to the users. This could include indicating the likelihood of cancer presence based on the model's output.

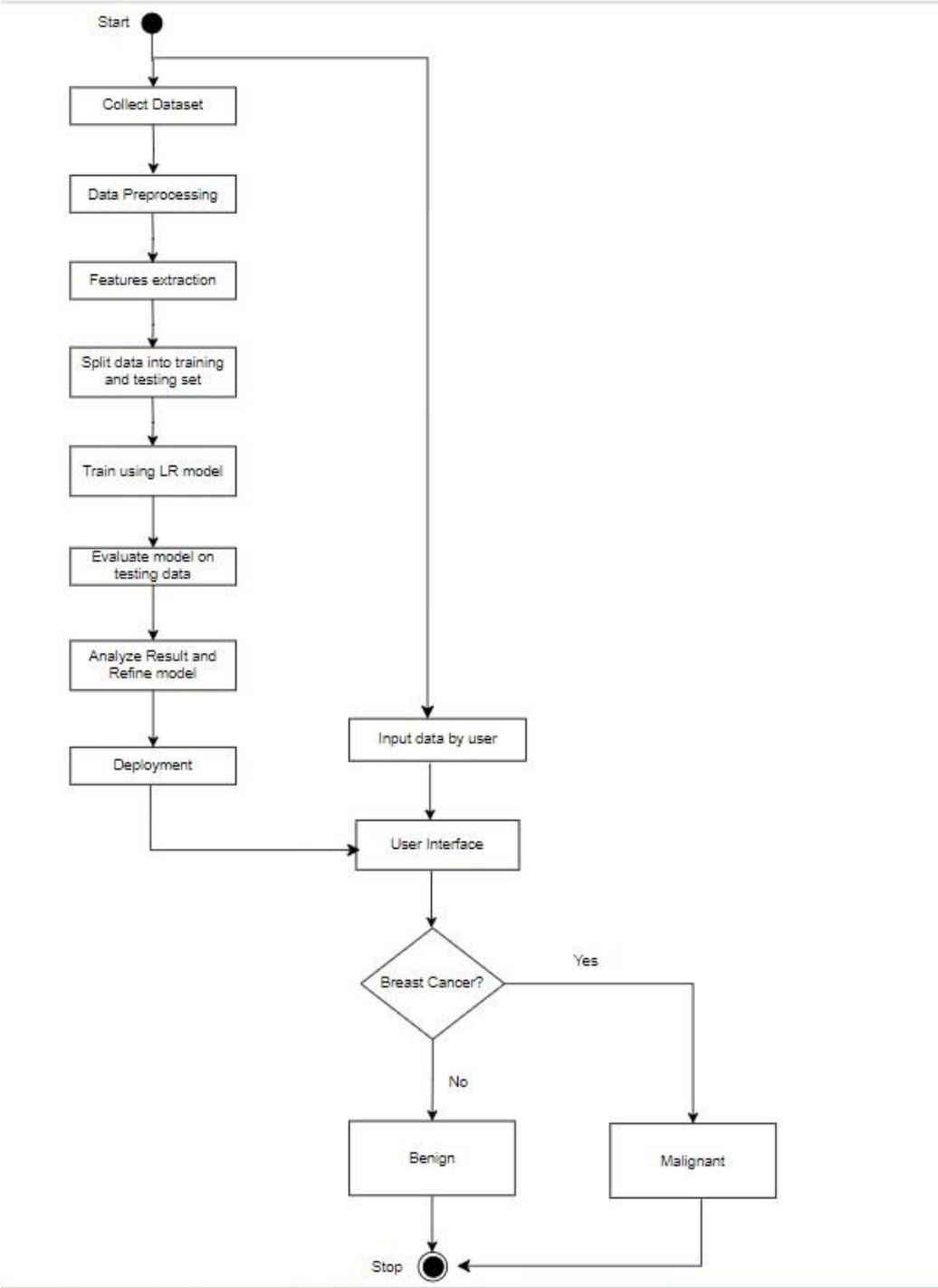


Figure 3.6: Activity Diagram for Breast Cancer Detection System

Chapter 4: System Design

4.1 Design

Here is a simple block diagram for “Breast Cancer Detection Using Logistic Regression” to show system workflow.

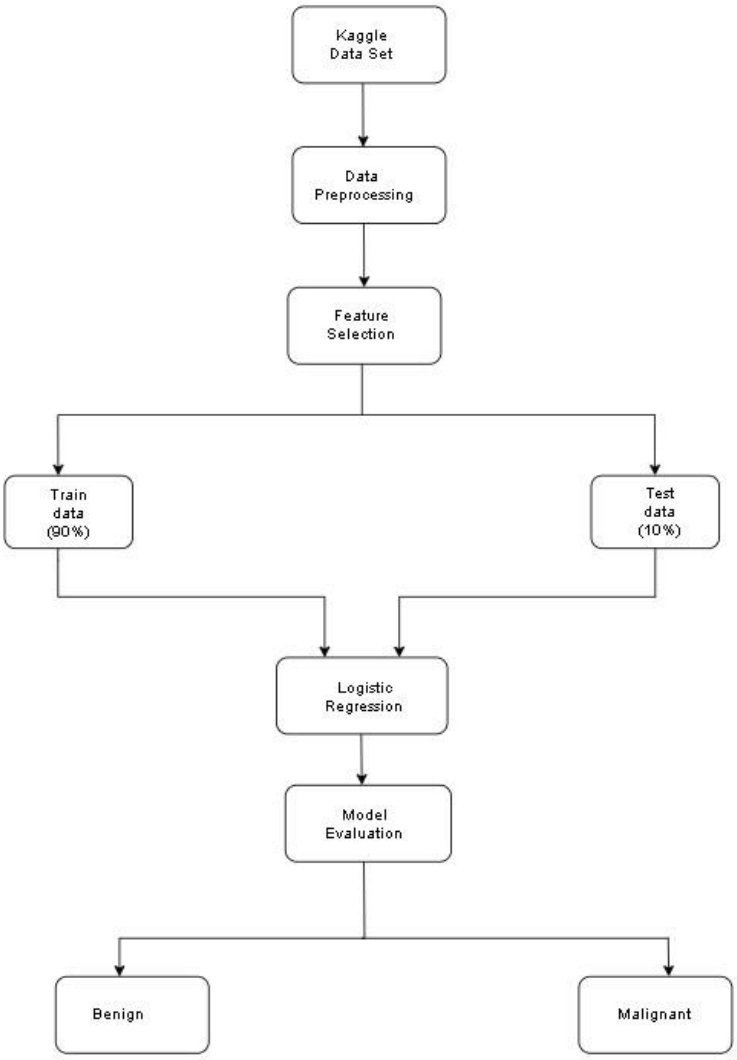


Figure 4.1: Block Diagram of Breast Cancer Detection System

4.1.2 Component Diagram

In depicts the general operation and other component of prediction process and represent a system physical component.

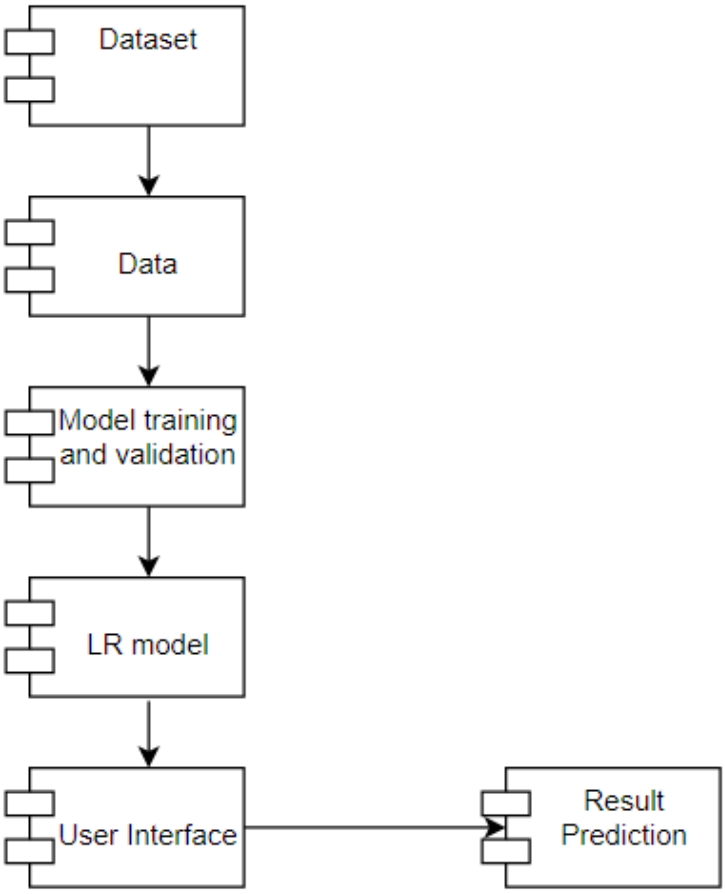


Figure 4.2: Component Diagram of Breast Cancer Detection System

4.2 Algorithms Details

The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1. This function is known as logistic function.

The odd is the ratio of something occurring to something not occurring. It is different from probability as the probability is the ratio of something occurring to everything that could possibly occur. So odd will be:

$$\frac{p(x)}{1-p(x)} = e^z$$

Applying natural log on odd. Then log odd will be:

$$\begin{aligned}
\log\left[\frac{p(x)}{1-p(x)}\right] &= z \\
\log\left[\frac{p(x)}{1-p(x)}\right] &= w \cdot X + b \\
\frac{p(x)}{1-p(x)} &= e^{w \cdot X + b} \quad \cdots \text{Exponentiate both sides} \\
p(x) &= e^{w \cdot X + b} \cdot (1 - p(x)) \\
p(x) &= e^{w \cdot X + b} - e^{w \cdot X + b} \cdot p(x)) \\
p(x) + e^{w \cdot X + b} \cdot p(x)) &= e^{w \cdot X + b} \\
p(x)(1 + e^{w \cdot X + b}) &= e^{w \cdot X + b} \\
p(x) &= \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}}
\end{aligned}$$

Then the final logistic regression equation will be:

$$p(X; b, w) = \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}} = \frac{1}{1 + e^{-w \cdot X + b}}$$

Likelihood function for Logistic Regression

The predicted probabilities will be:

For y=1 The predicted probabilities will be: p(X:b,w)=p(x)

For y =0 The predicted probabilities will be: 1-p(X:b,w)=1-p(x)

$$L(b,w) = \prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$$

Taking natural logs on both sides

$$\begin{aligned}
\log(L(b,w)) &= \sum_{i=1}^{n} y_i \log p(x_i) \; + \; (1-y_i)\log(1-p(x_i)) \\
&= \sum_{i=1}^{n} y_i \log p(x_i) + \log(1-p(x_i)) - y_i \log(1-p(x_i)) \\
&= \sum_{i=1}^{n} \log(1-p(x_i)) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1-p(x_i} \\
&= \sum_{i=1}^{n} -\log 1 - e^{-(w\cdot x_i+b)} + \sum_{i=1}^{n} y_i(w\cdot x_i+b) \\
&= \sum_{i=1}^{n} -\log 1 + e^{w\cdot x_i+b} + \sum_{i=1}^{n} y_i(w\cdot x_i+b)
\end{aligned}$$

Gradient of the log-likelihood function

To find the maximum likelihood estimates, we differentiate with respect to w.

$$\begin{aligned}
\frac{\partial J(l(b,w)}{\partial w_j} &= -\sum_{i=n}^{n} \frac{1}{1+e^{w\cdot x_i+b}} e^{w\cdot x_i+b} x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \\
&= -\sum_{i=n}^{n} p(x_i;b,w)x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \\
&= \sum_{i=n}^{n} (y_i - p(x_i;b,w))x_{ij}
\end{aligned}$$

Chapter 5: Implementation and Testing

5.1 Implementation

Implementation refers to the process of putting plans, strategic, policies or projects into action or practice. It involves taking concrete steps and executing the planned activities to achieve the intended goals or objectives. First, we collected the data from the kaggle dataset. Then, we performed data pre-processing using various techniques. The preprocessed data was split into training, validation and testing sets. The model was trained for 569 dataset. The model's performance was evaluated using the testing dataset and appropriate evaluation metrics using different attributes of breast cancer dataset.

5.2 Tools Used

i.Programming Language

Python is a high-level, interpreted programming language that is widely used for a variety of purposes, including web development, scientific computing, data analysis and machine learning. Its popularity can be attributed to its simplicity, readability and versatility, making it an excellent choice for both beginner and experienced programmers alike. With an extensive library of modules and a thriving community of developers, python is an excellent tools for solving complex problem. Fundamental python libraries like numpy, pandas, sklearn,matplot,pillow etc are used in this study.

Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of web development, so we can focus on writing app without needing to reinvent the wheel. It's free and open Source. So we use this framework in this study.

ii.Integrated Development Environment

A web-based interactive computational environment VS code was used to write a python codes. This was used under the installation of python 3.12.0 version.

iii. Documentation Tools

- Ms-Word was used as a text editor for documentation process.
- Draw.io was used to draw different diagram used in this system.

5.3 Testing

The testing phase depicts the efficiency and inefficiency of the system.

5.3.1 Unit Testing

Unit testing refers to the testing of every small modular component of the system , keeping them isolated from other modules. Here we mention testing result of the various part of the system. In unit testing, we design the whole system in modularized pattern and each module was tested. Till we get the accurate output from the individual module we have worked on the same module. We have checked for the outcome of each module. We use every unitary module of our system.

i.Test case 1: Submit data

ii.Test case 2: Detection of cancer

Case 1	Expected Result	Actual Result	Conclusion
Submit new data	User should able to submit the data	Data Submitted Successfully	Successful

Table 5.1: Test case for Submitting Data

Case 2	Detection of the cancer based on data provided.
Expected Result	System should detect the cancer from the data provided and redirect to result page.
Actual result	Breast cancer can either be benign or malignant is predicted and redirect to the result page.
Conclusion	Successfully

Table 5.2: Test Case for Detecting Cancer

5.3.2 System Testing

In this testing phase, our system was tested. Every individual component was integrated and tested against user and hardware compatibility. Sometime in this testing process dependency error was found due to different local server environment and dependency conflict .To overcome this problem virtual environment were used.

S.N	Test Description	Expected Result	Actual Result	Remarks
1	Test data benign (1)	Benign	Benign	Pass
2	Test data Malignant (1)	Malignant	Malignant	Pass
3	Test data Malignant (4)	Malignant	Benign	Fail
4	Test data Benign (2)	Benign	Benign	Pass
5	Test data Malignant (2)	Malignant	Malignant	Pass

Table 5.3: System Testing

5.4 Result Analysis

5.4.1 Feature value in Index 0

During the study test record of index 0 was fed to have the explainable having details as shown in table 5.

Features	Values
Radius mean	17.99
Texture mean	10.38
Perimeter mean	122.8
Area mean	1001
Smoothness mean	0.1184
Compactness mean	0.2776
Concavity mean	0.1471
Concave point mean	0.1471
Symmetry mean	0.2419
Fractal dimension mean	0.07871
Radius se	1.095
Texture se	0.9053
Perimeter se	8.589
Area se	153.4
Smoothness se	0.006399
Compactness se	0.04904
Concavity se	0.05373
Concave point se	0.01587
Symmetry se	0.03003
Fractal dimension se	0.006193
Radius worst	25.18
Texture worst	17.33
Perimeter worst	184.6
Area worst	2019

Smoothness worst	0.1622
Compactness worst	0.6656
Concave point worst	0.2654
Symmetry worst	0.4601
Fractal dimension worst	0.1189

Table 5.4: Feature Values in Index 0

5.4.2 Confusion Matrix

Figure 5.1 shows that confusion matrix for the prediction of breast cancer. Out of total 187 Benign samples only 3 samples were predicted wrong. On the other hand, only 2 samples of Malignant were mis-classified where 96 samples were predicted true.

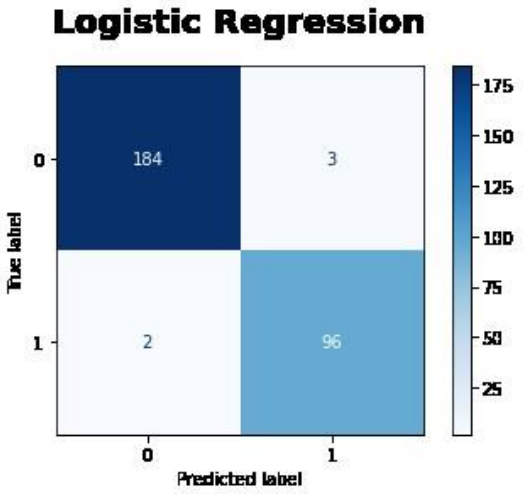


Figure 5.1: Confusion Matrix

Here's a brief explanation of a confusion matrix and its components in a binary classification scenario (e.g., malignant or benign):

a) True Positive (TP): Instances where the model correctly predicts malignant cases.
b) True Negative (TN): Instances where the model correctly predicts benign cases.
c) False Positive (FP): Instances where the model incorrectly predicts malignant cases when they are actually benign (Type I error).

d) False Negative (FN): Instances where the model incorrectly predicts benign cases when they are actually malignant (Type II error).

5.4.3 AUC Curve

Figure 13 shows AUC train values as 98.39% and test value as 98.17%. Result value show that a LR model is performing well in distinguishing between positive and negative samples.

+

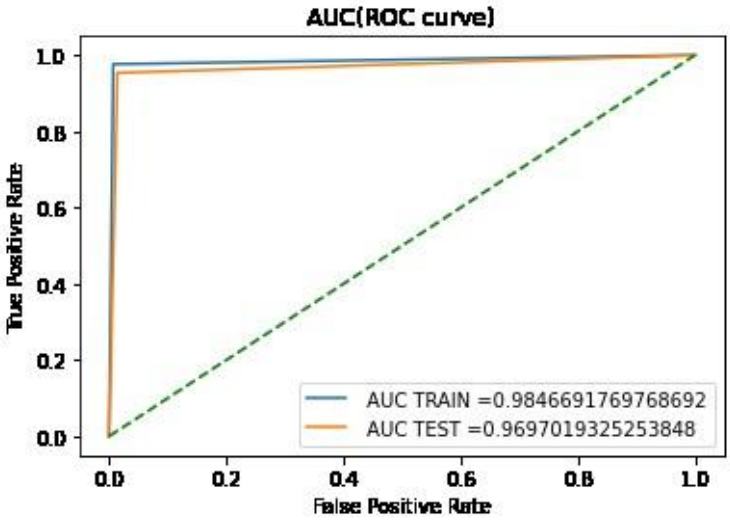


Figure 5.2: LR AUC Curve

Chapter 6: Conclusion and Future Recommendation

6.1 Conclusion

The purpose of creating the “Breast Cancer Detection Using Logistic Regression” project was to automatically identify cancerous tissue on dataset. This was achieved by using a Logistic Regression to classify the data into the different categories and training it on the kaggle dataset.

After training and evaluating the model, the result showed that the architecture achieved an accuracy of 85% on the test set. The project also analyzed the training and validation accuracy. This project has the potential to assist medical professionals in accurately diagnosing breast cancer at an earlier stage, leading to more effective treatment and improved patients outcomes.

Overall, the project successfully utilized Logistic Regression to classify the data attributes including the presence or absence of cancerous tissue in the breast. Further research and refinement of this technology could contribute to the development of more accurate and efficient methods for detecting breast cancer.

6.2 Future Recommendation

“Here are some future recommendations for the “Breast Cancer Detection Using LR” project:

Increase the dataset size: The performance of a LR model greatly depends on the amount and quality of data used to train it. Collecting more data and using data augmentation techniques can help improve the model’s accuracy.

Data Augmentation: Data augmentation techniques can be used to increase the size of the dataset, which will help the model learn more robust features. Techniques such as random rotations, translations and flips can be used to generate new training data.

Explore different architectures: Try experimenting with different LR model and consider including the patient data, such as age, family history, and lifestyle factors, to see if they can help improve the accuracy of the model.

References

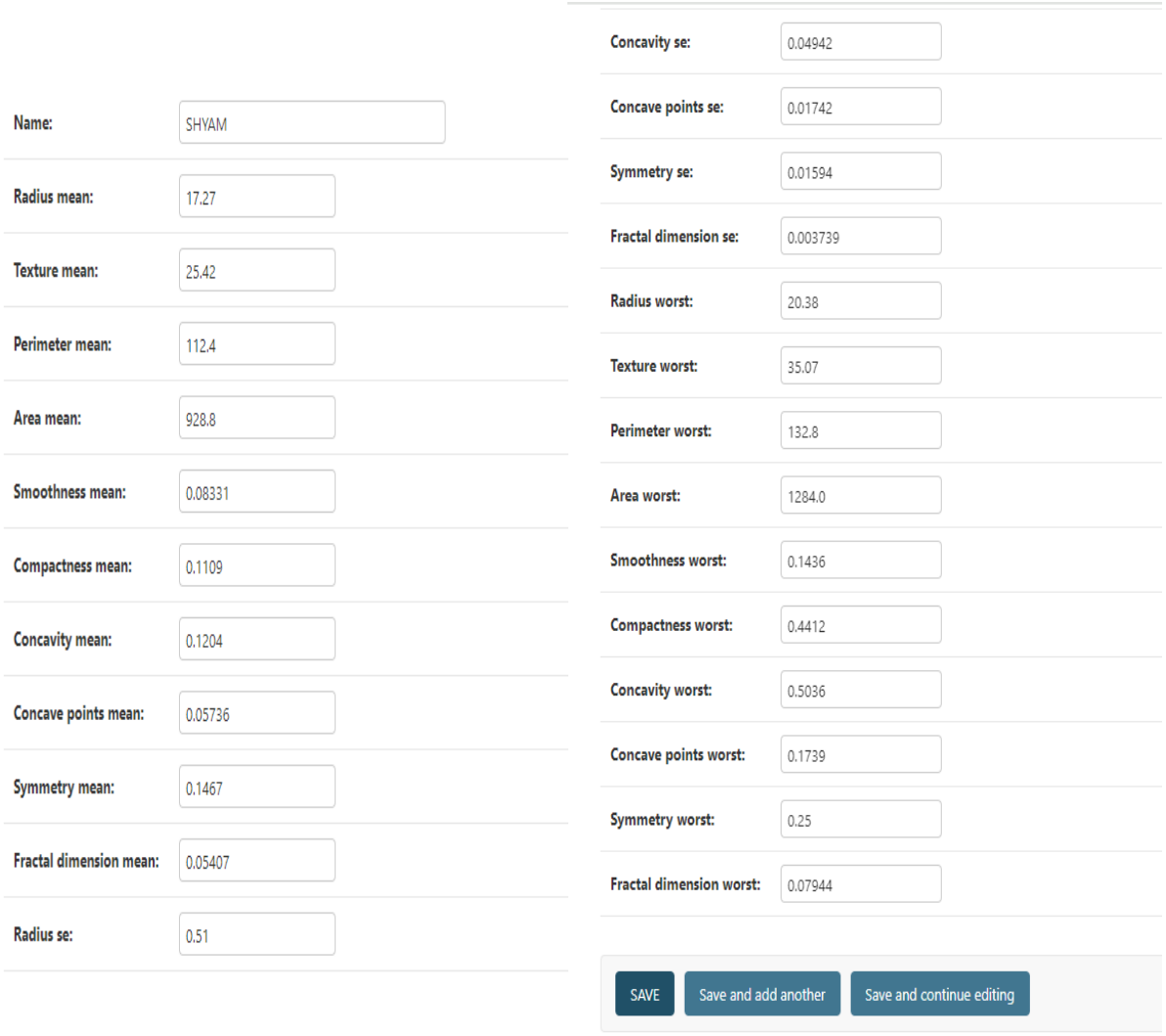
[1] Aruna, S.; Rajagopalan, S.; Nandakishore, L. Knowledge based analysis of various statistical tools in detecting breast cancer. Comput. Sci. Inf. Technol. 2011, 2, 37–45.

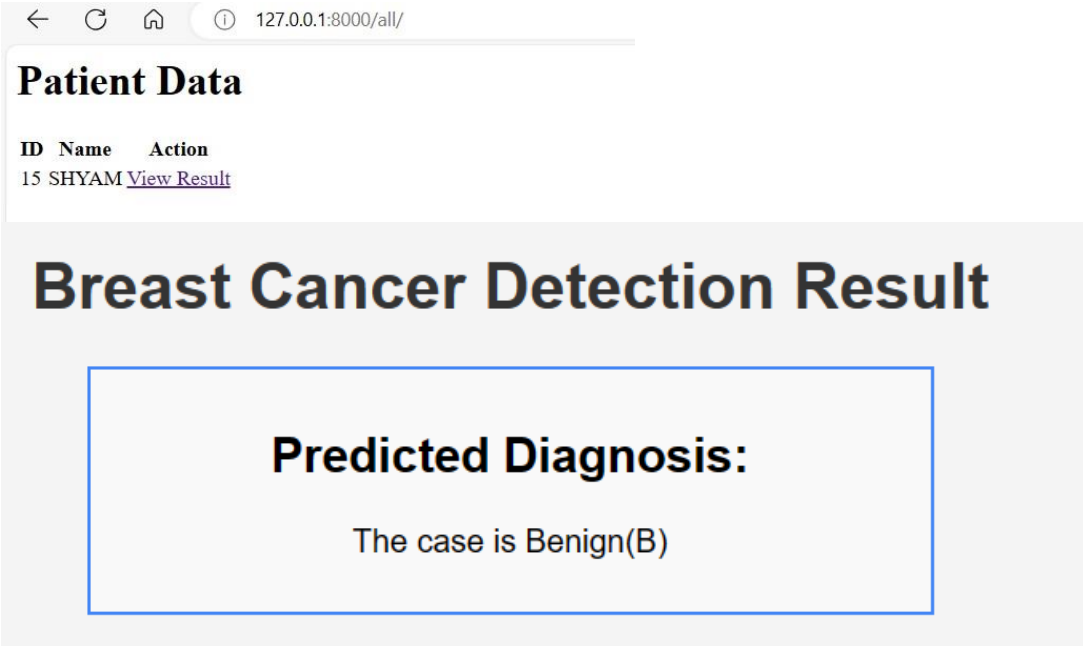
[2] Chaurasia, V.; Pal, S. Data mining techniques: To predict and resolve breast cancer survivability. Int. J.Comput. Sci. Mob. Comput. 2014, 3, 10–22.

[3] Asri, H.; Mousannif, H.; Al Moatassime, H.; Noel, T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput. Sci. 2016, 83, 1064–1069.

[4] Bernal, J.L.; Cummins, S.; Gasparrini, A. Interrupted time series regression for the evaluation of public health interventions: A tutorial. Int. J. Epidemiol. 2017, 46, 348–355.

Annex 1: Snapshots





Prediction Result

Data Train Code

```
def standardize_data(X):
    mean = np.mean(X, axis=0)
    std = np.std(X, axis=0)
    standardized_X = (X - mean) / std
    return standardized_X

# first Standardize the attributes of the breast dataset and we have to drop
diagnosis column
X = data.drop(columns=['diagnosis'])
X_standardized = standardize_data(X)


# use sigmoid function
def sigmoid(z):
    return 1 / (1 + np.exp(-z))

def logistic_regression(X, y, learning_rate=0.01, num_iterations=1000):
    num_samples, num_features = X.shape
    """_summary_

    Returns:
        _type_: _description_
    """
    theta = np.zeros(num_features)

    for _ in range(num_iterations):
        # Calculate the hypothesis and the gradient
        z = np.dot(X, theta)
        h = sigmoid(z)
        gradient = np.dot(X.T, (h - y)) / num_samples

        # Update the parameters (theta)
```

```
        theta -= learning_rate * gradient

    return theta

# Add bias term to the input features
X_bias = np.c_[np.ones((X_standardized.shape[0], 1)), X_standardized]

# Convert diagnosis labels to binary (0: Benign, 1: Malignant)
y = data['diagnosis'].map({'M': 1, 'B': 0}).values

# Train the logistic regression model
coefficients = logistic_regression(X_bias, y)

def predict(X, coefficients):
    X_bias = np.c_[np.ones((X.shape[0], 1)), X]
    z = np.dot(X_bias, coefficients)
    return sigmoid(z)

def logistic_regression_model(X, coefficients):
    X_bias = np.c_[np.ones((X.shape[0], 1)), X]
    z = np.dot(X_bias, coefficients)
    return sigmoid(z)

def predict_diagnosis(request,id):
    data = BreastCancerData.objects.get(id=id)

    input_data = {
        'id':data.id,
        'radius_mean': data.radius_mean,
        'texture_mean':data.texture_mean,
        'perimeter_mean': data.perimeter_mean,
         'area_mean':data.area_mean,
         'smoothness_mean':data.smoothness_mean,
        'compactness_mean':data.compactness_mean,
        'concavity_mean':data.concavity_mean,
```

```
        'concave_points_mean':data.concave_points_mean,
        'symmetry_mean':data.symmetry_mean,
        'fractal_dimension_mean':data.fractal_dimension_mean,
        'radius_se':data.radius_se,
        'texture_se':data.texture_se,
        'perimeter_se':data.perimeter_se,
        'area_se':data.area_se,
        'smoothness_se':data.smoothness_se,
        'compactness_se':data.compactness_se,
        'concavity_se':data.concavity_se,
        'concave_points_se':data.concave_points_se,
        'symmetry_se':data.symmetry_se,
        'fractal_dimension_se':data.fractal_dimension_se,
        'radius_worst':data.radius_worst,
        'texture_worst':data.texture_worst,
        'perimeter_worst':data.perimeter_worst,
        'area_worst':data.area_worst,
        'smoothness_worst':data.smoothness_worst,
        'compactness_worst':data.compactness_worst,
        'concavity_worst':data.concavity_worst,
        'concave_points_worst':data.concave_points_worst,
        'symmetry_worst':data.symmetry_worst,
        'fractal_dimension_worst':data.fractal_dimension_worst
        # Add other attributes for your input data
    }
    # Create a DataFrame from the input data
    input_df = pd.DataFrame([input_data])

    # Standardize the input data using the same mean and std as the training
data
    input_standardized = standardize_data(input_df)

    # Get the logistic regression prediction
    prediction = predict(input_standardized, coefficients)[0]
    print(f"Prediction: {prediction}")
```

```
    # Convert the prediction to a diagnosis label
 # Convert the prediction to a diagnosis label
    predicted_diagnosis = 'The data you have entered is Malignant(M)' if
prediction >= 0.5 else  'The data you have entered is Benign(B)'
```